140012065 Chris Schwaab 19th April 2016

Overview:

The objective of this project was to gain experience using Python to work with large data sets. The practical aimed to improve our programming skills and get more insights into the projects from the Python ecosystem such as libraries for data analysis and visualisation and Jupyter notebooks.

Requirements Fulfilled:

The below show all the requirements that the team fulfilled

Basic:

- refine the dataset
- perform the descriptive analysis of the dataset by:

  – calculating the total number of tweets, retweets and replies in the dataset
  – calculating means and standard deviations for the number of tweets, retweets and replies sent by each user
  – identifying most popular hashtags

- build plots/visualisations for:

  – the structure of the dataset (tweets/retweets/replies) – the timeline of the tweets activity
  – the hashtag cloud

- provide the Jupyter notebook to re-run the analysis (starting from the raw or refined data).

Extensions:

- **Easy:** Analyse applications used to send tweets.
- **Easy:** Extend the descriptive analysis, for example, by calculating means and standard deviations for the number of times each user being retweeted and the number of times each user being replied.
- **Medium:** Analyse the pattern of users' activity over the period covered by the dataset.
- **Medium to Hard:** Analysing interactions between users by constructing and analysing the graph based on retweets, replies and mentions. (50% finished).
- **Hard:** Implementing interactive visualisiations.
- **Hard:** Analysing some other, possibly much larger, data sets.

Design:

The python project has the following files that are used to make the analysis. There also exist .html and json files in order to create interactive visualisations.

.py files:

- parse.py:
    - This file contains the methods in which the csv file is refined. The way it is refined is that it creates a new data frame from which pandas can work on. It gives 'source' a new column in the data frame. It loads the json in entities_str to the data frame and refines the created_at time field in the set.
- analyse.py
    - This file contains all the analysis needed to analyse the given data set. All the analysis methods return data frames which then make it easier to convert to a json file hence making the interactive graphs easier to do.
- encode.py
    - Encodes the data frame into a json file
- test.py
    - Used to test methods and debugging
- graph.py
    - Not used anymore, however it creates the hashtag cloud.

.html files: (To make the interactive graphs we used the http://d3js.org library)

- appsToSendTweets.html: makes bar graph that charts number of tweets sent by applications.
- popularMentions.html: makes cloud that gets most popular user.
- tweetBubble.html: makes hashtag cloud.
- tweets_per_time.html: makes timeline of the different types of tweets
- types_pieChart.html: makes pie chart of the sum of all the types. The structure of the data set.
- retweetsBubble.html:
- replyBubble.html:

.json files:

- Apps_count.json
- displayData.json
- hashtags_count.json
- mentions.json
- replies_per_user.json
- retweets_per_user.json
- sum_types.json
- tweet_time.json
- Types.json

Testing:
The jupyter notebook shows the tests that were done. Please check the pdf document python_notebook_core to see some of the results or open the python notebook.

Implementation:
This part shows the implementations that I done and my contributions to the group work at hand.

Basic:

- refine the dataset 140013444
- perform the descriptive analysis of the dataset 140013444:
  – calculating the total number of tweets, retweets and replies in the dataset
  – calculating means and standard deviations for the number of tweets, retweets and replies sent by each user 140013444
  – identifying most popular hashtags with 140013444

- build plots/visualisations for:
  – the structure of the dataset (tweets/retweets/replies) – the timeline of the tweets activity done by me.
  – the hashtag cloud: done by Peter

- provide the Jupyter notebook to re-run the analysis (starting from the raw or refined data): Done by me

Extensions:

- **Easy:** Analyse applications used to send tweets: Done by me
- **Easy:** Extend the descriptive analysis, for example, by calculating means and standard deviations for the number of times each user being retweeted and the number of times each user being replied. Done by me
- **Medium:** Analyse the pattern of users' activity over the period covered by the dataset. Done by me.
- **Medium to Hard:** Analysing interactions between users by constructing and analysing the graph based on retweets, replies and mentions. (50% finished). Done by me. Did not have a network graph. Instead had three cloud graphs.
- **Hard:** Implementing interactive visualizations.

  - appsToSendTweets.html: Done by me
  - popularMentions.html: Done by me
  - tweetBubble.html: done by Peter
  - tweets_per_time.html: Done by me and with of 140013444
  - types_pieChart.html: Done by me
  - retweetsBubble.html: Done by me
  - replyBubble.html: Done by me

- **Hard:** Analysing some other, possibly much larger, data sets. Done by me.

Evaluation:

I believe that the requirements of the practical have been met with very few hindrances. The biggest problem would be not being able to get the network graph working properly. I think that the team worked well together with some hiccups along the way.

Conclusion:

In Conclusion I feel that I have been able to learn much about python. I also think that my analysis techniques have improved due to this practical. I also feel that I am now more comfortable working in a team. I also have a greater understanding of the different libraries python uses.